



From Transactions to Tokens: Building the AI Factory for Financial Services

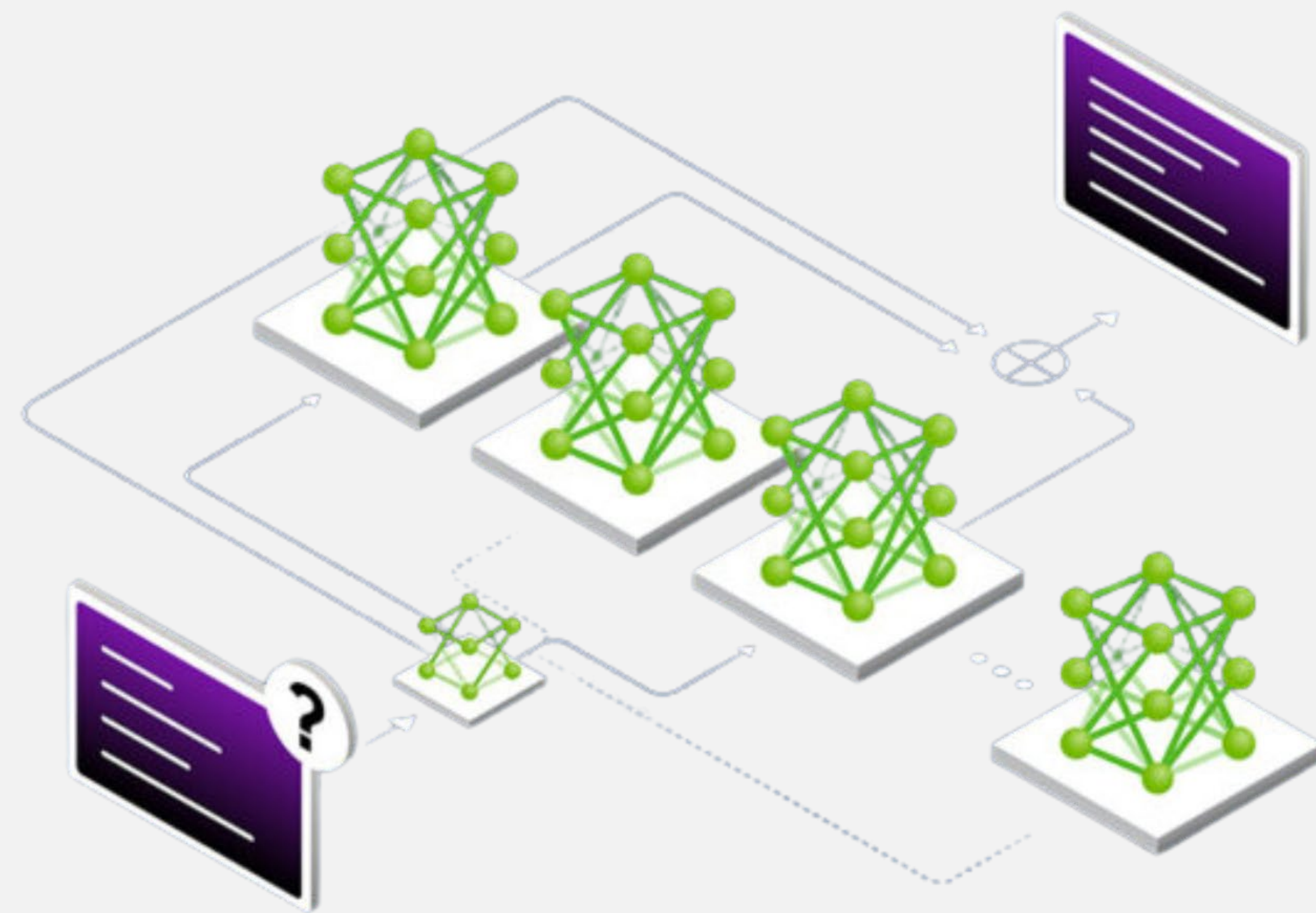
Marc Stampfli, Business Director Switzerland

E-Mail: mstampfli@nvidia.com

Linkedin: <https://www.linkedin.com/in/marcstampfli>



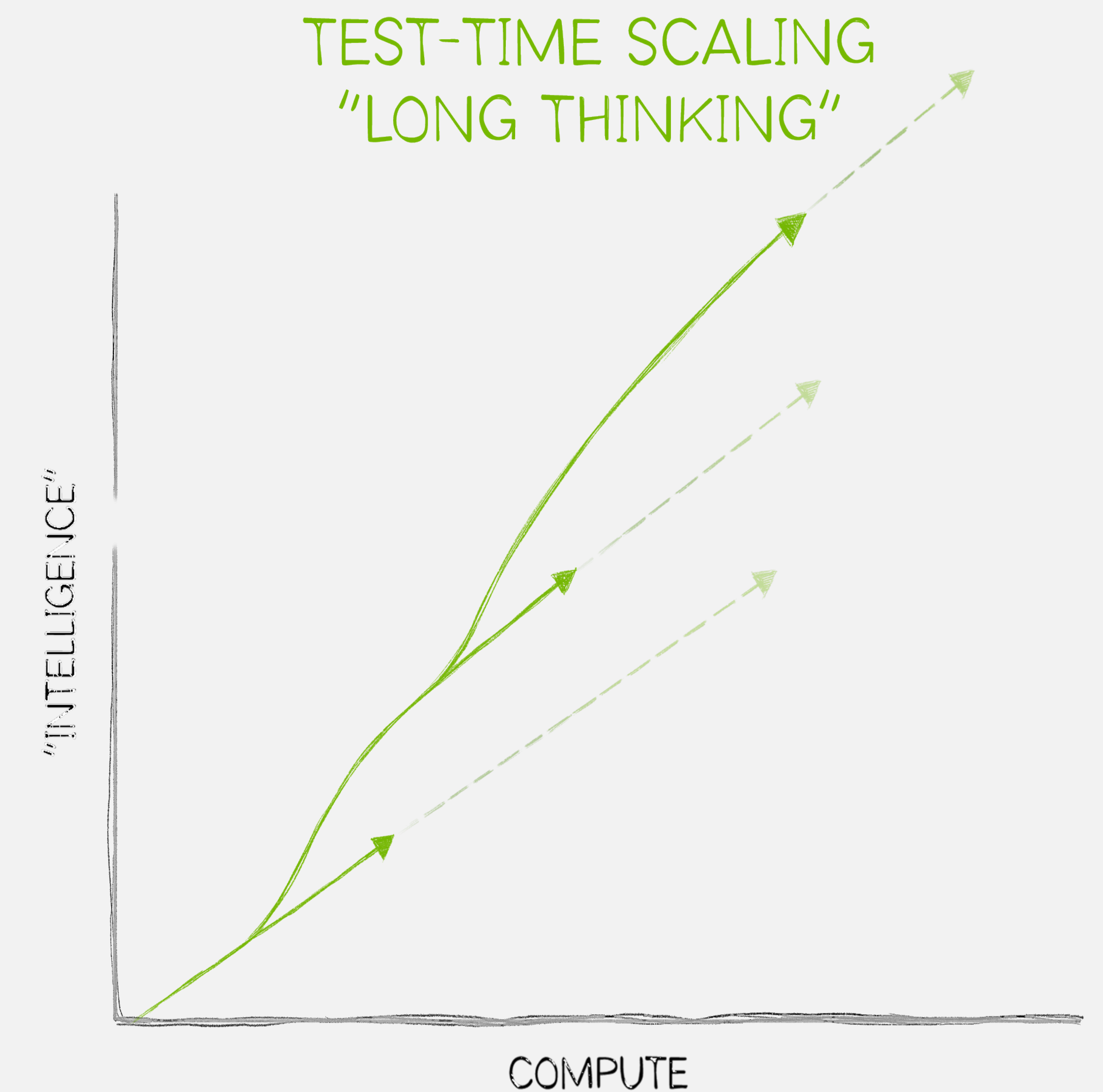
What Long-Running AI Agentic Systems Need



System of Models



Specialized Agents



High Efficiency

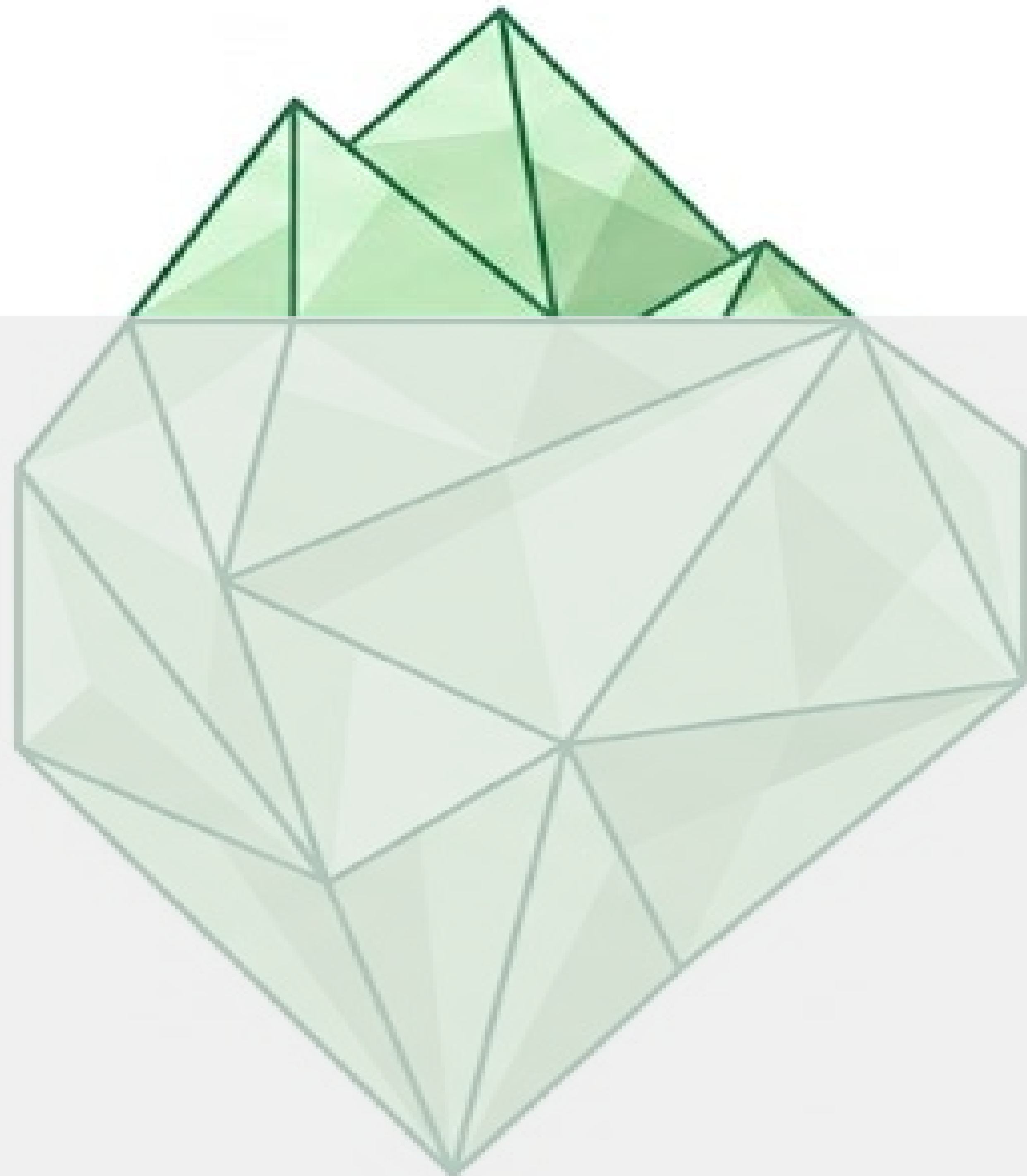
AI Factories are the Industrial Infrastructure of the AI Era

Inference is the Workload
Tokens are the New Commodity
Compute is Revenue



Inference Iceberg: Are you looking at the right metrics?

Optimizing for input when business runs on output is a fundamental mismatch



Peak Chip Specifications such as FLOPS and HBM
Cost per GPU per hour
FLOPs per dollar

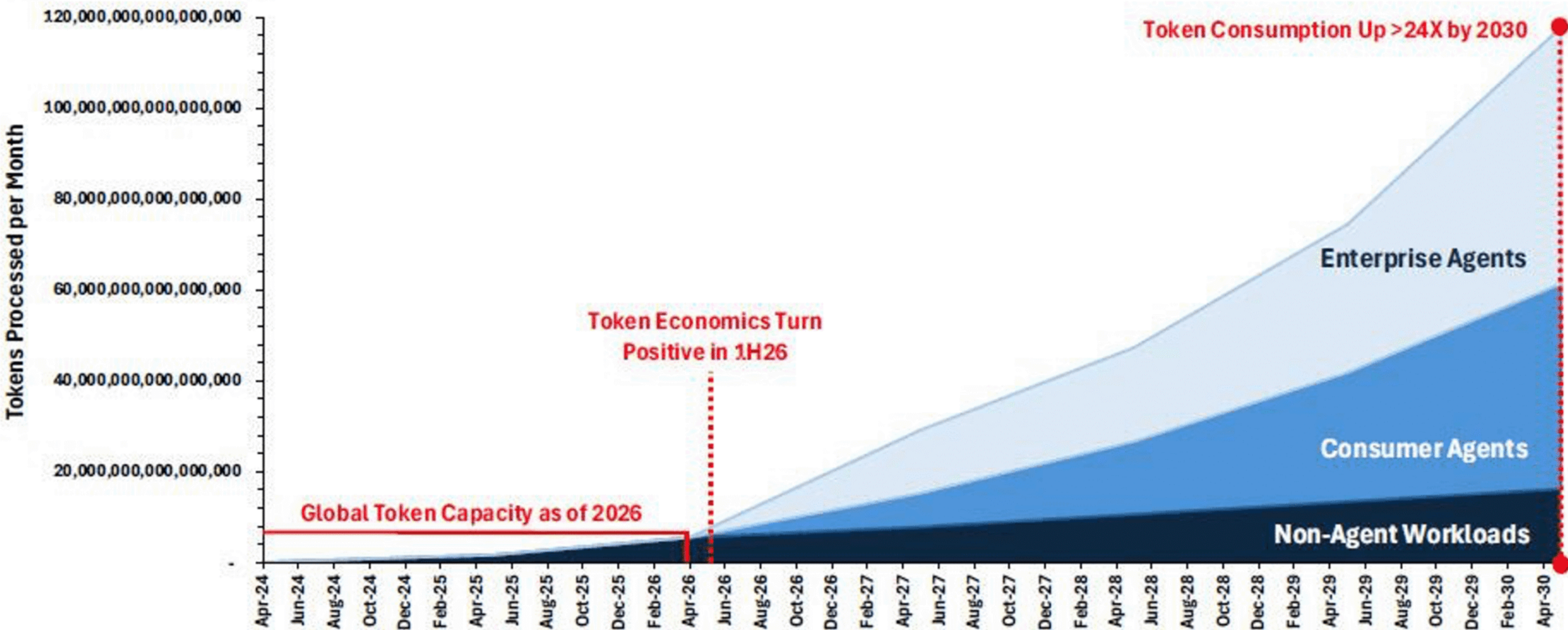
Extreme Co-design across compute, networking, memory, storage, software, and ecosystem

$$\text{Cost per M tokens} = \left[\frac{\text{Cost per GPU per hour}}{\text{Tokens per GPU per Hour}} \right] \times 1 \text{ million}$$

*Tokens per GPU per Hour = Tokens/sec x 3600

Technology and R&D Budgets are Becoming Token Budgets

Exhibit 2: By 2030, we estimate consumer and enterprise agents could push token consumption 24X above today's estimated global capacity.



Source: Data compiled by Goldman Sachs Global Investment Research

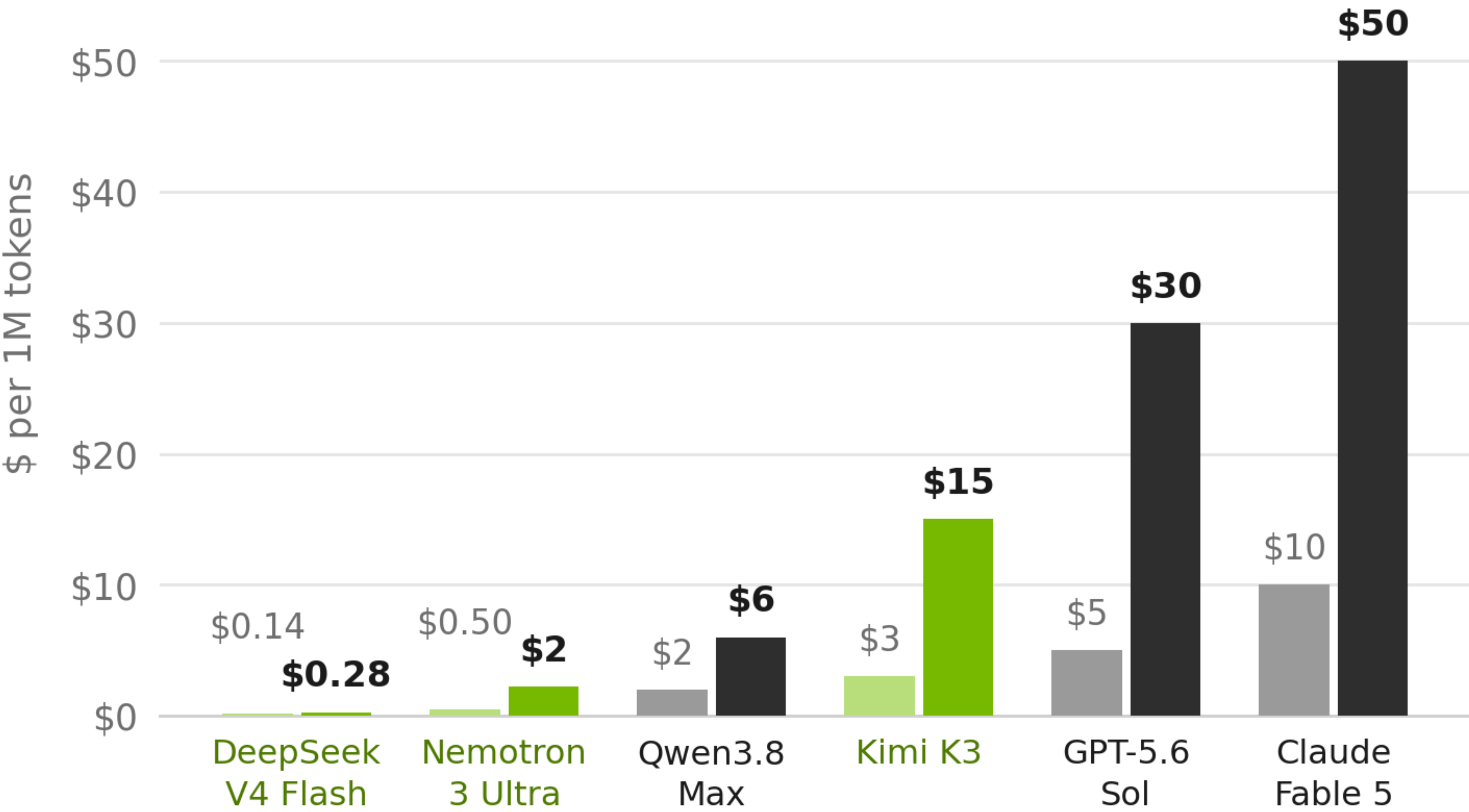
Price and capability across six frontier models

Published list rates and the Artificial Analysis Intelligence Index v4.1 · August 2026

● Closed weights ● Open weights

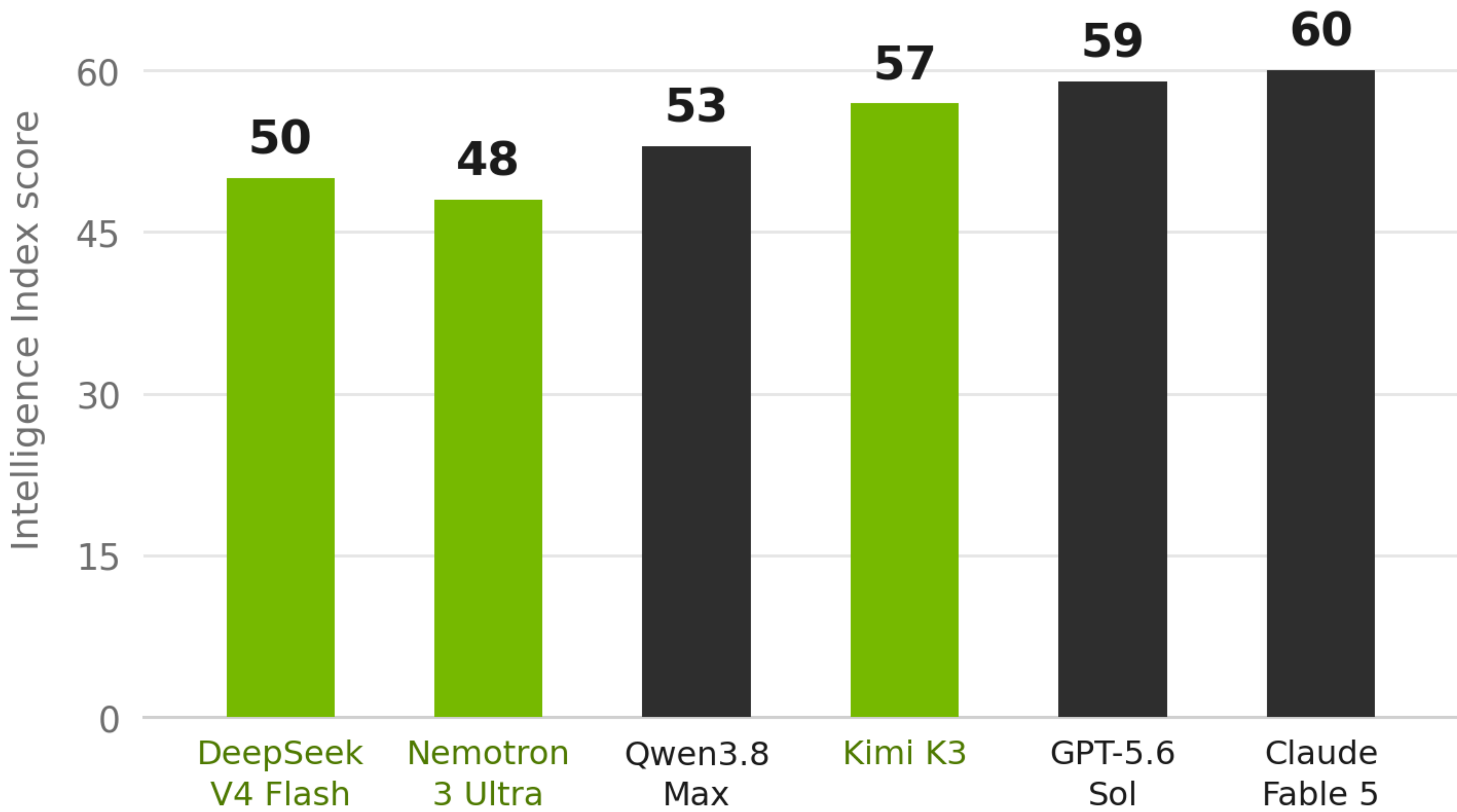
What you pay

Input and output list price, \$ per 1M tokens



What you get

Artificial Analysis Intelligence Index v4.1



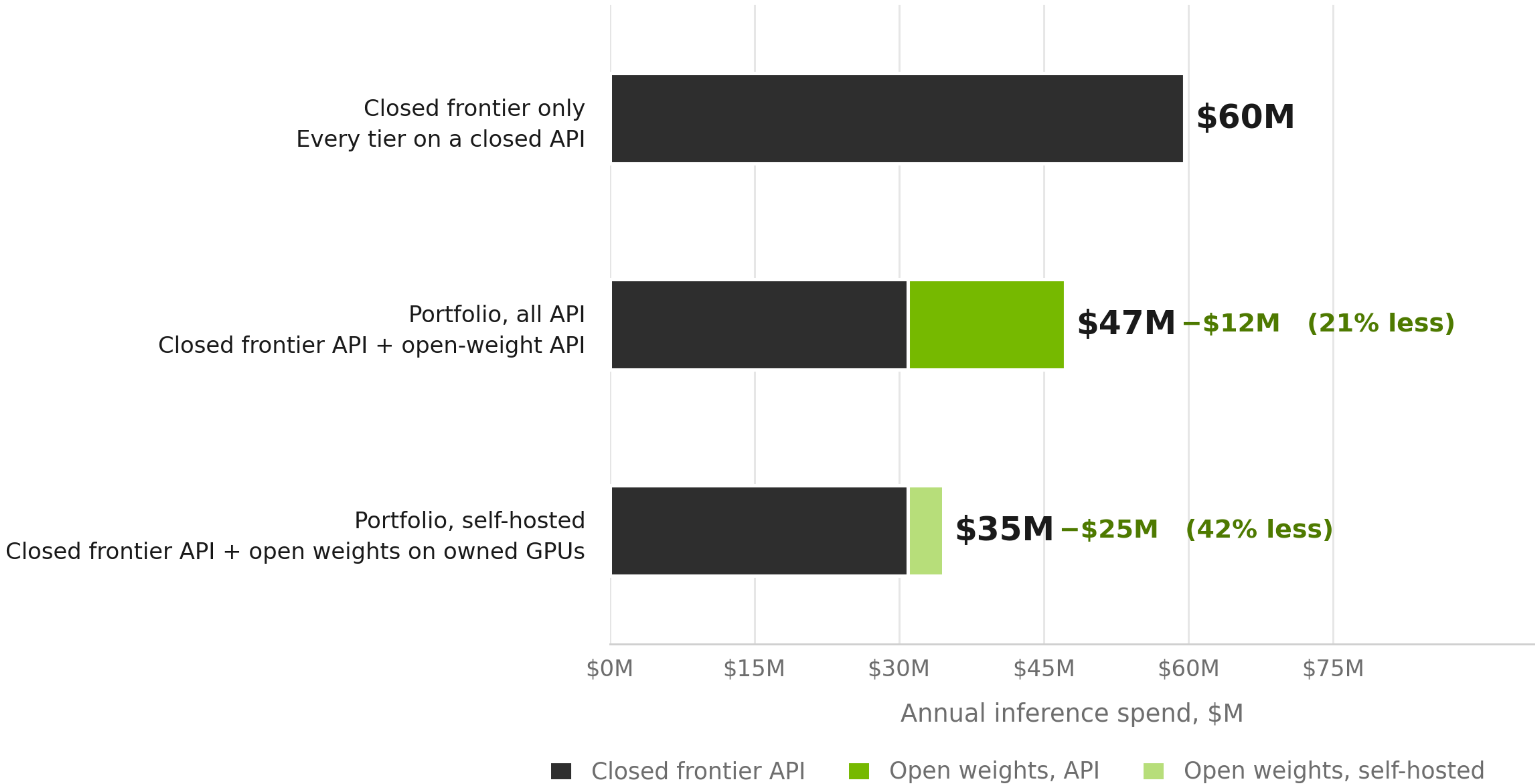
From DeepSeek V4 Flash 0731 to Claude Fable 5, list price rises 71x on input and 179x on output for 10 points of index.

Nemotron 3 Ultra is the only model shipping a fully open stack: weights, data, recipes, eval tooling and RL infrastructure under OpenMDW.

Sources: first-party rate cards — Anthropic, OpenAI, Moonshot AI, Alibaba Model Studio, DeepSeek (accessed 4 Aug 2026); Nemotron 3 Ultra at OpenRouter reference rates, as an open-weight model with no first-party per-token card (Fireworks \$0.60/\$2.40; a free route is also available); Artificial Analysis Intelligence Index v4.1 (artificialanalysis.ai/models). List prices exclude caching and batch discounts, which materially lower realized cost — DeepSeek’s first-party cache-hit rate is \$0.0028/M, a 98% discount. Qwen3.8-Max is closed-weight at time of writing; DeepSeek V4 Flash 0731 weights are pending release.

Financial Implications of Different Model Deployments

Annual inference spend at ~42,000 employees and 20T tokens/year · all three strategies clear the same intelligence floor on every workload



-\$12M
by adding open weights to the mix

Open models take tier 1 and the bulk tiers;
Opus 5 and Grok 4.5 keep the hardest work.

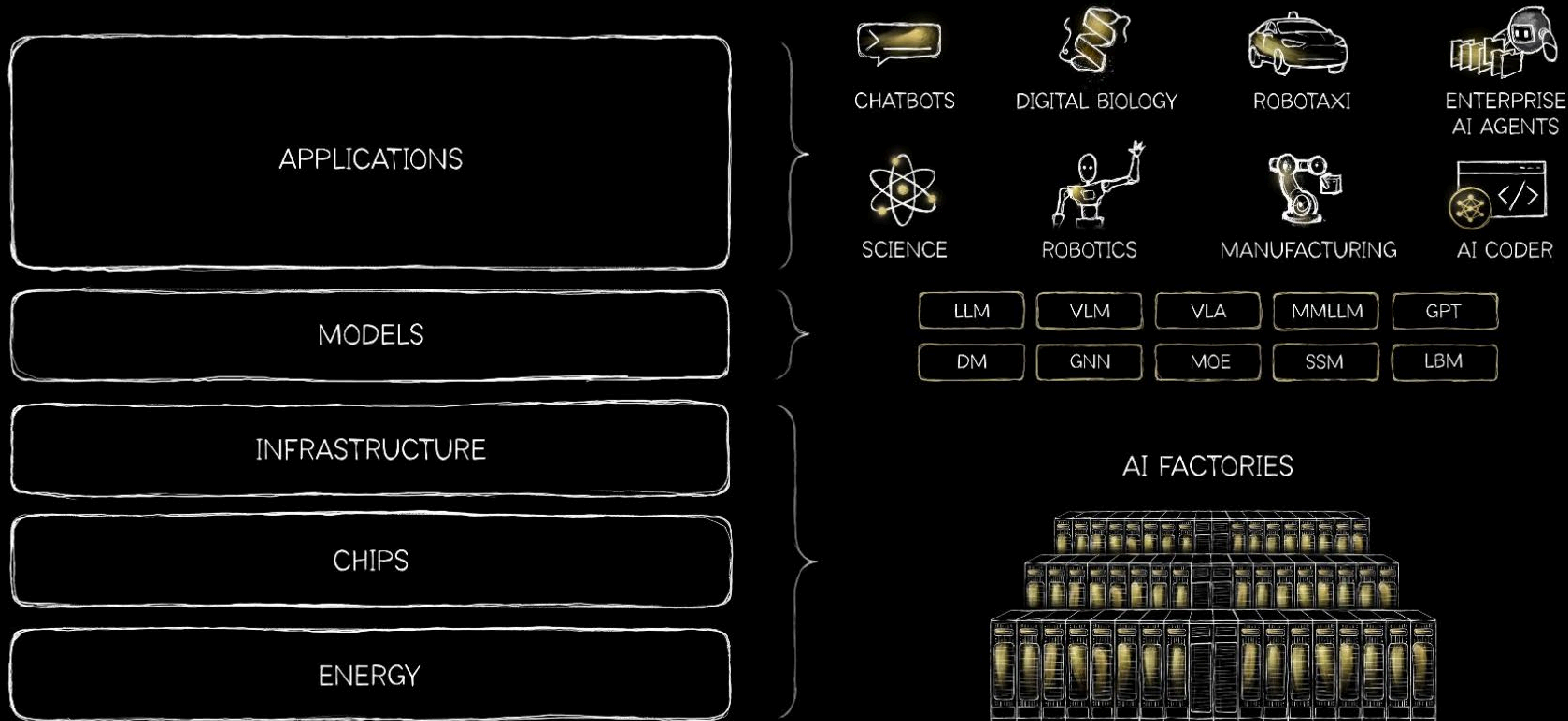
-\$25M
by self-hosting the open half

Same models, owned capacity at ≥70%
utilization. Break-even is ~5M tokens/day.

Model: capability tiers are floors on the Artificial Analysis Intelligence Index v4.1 (T0 best available, T1 ≥57, T2 ≥53, T3 ≥48, T4 ≥43); each strategy routes every tier to its cheapest qualifying model, so delivered capability is identical across all three. Volumes: 20,000 agentic-coding seats at 60M tokens/month (Anthropic reports \$150–250/dev/month enterprise Claude Code spend; this model yields \$210) and 30,000 knowledge-work seats at 8M/month. API prices per slide 1. Self-hosted rates (\$1.00/\$0.15/\$0.07 per M) are modeled, not quoted: reserved capacity at ≥70% utilization, 3-year amortization, power and ops included. Excludes fine-tuning, evals, and platform headcount (budget 3–6 FTE).

AI is a Five-Layer Cake

AI as a layered infrastructure that optimizes cost, control, and intelligence



Build your own Transaction Foundation Model (TFM)

Revolut



FSI Developer Example — Build Your Own Transaction Foundation Model

v:2604.08649v1 [cs.LG] 9 Apr 2026

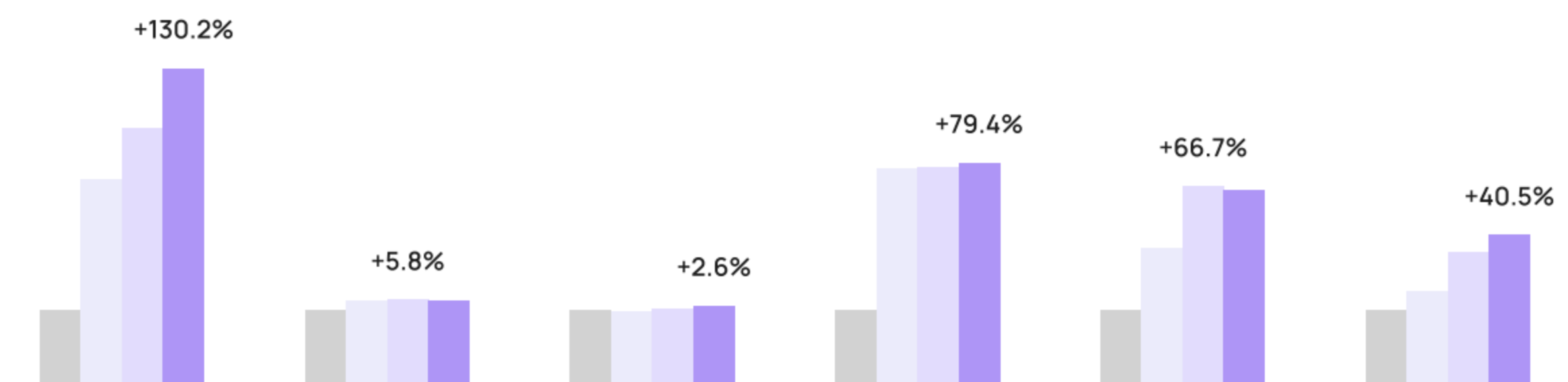
PRAGMA: Revolut Foundation Model

Maxim Ostroukhov¹ Ruslan Mikhailov¹ Vladimir Iashin¹
Artem Sokolov¹ Andrei Akshonov¹ Vitaly Protasov¹ Dmitrii Beloborodov¹
Vince Mullin² Roman Y. Enzmann² Georgios Kolovos² Jason Renders²
Pavel Nesterov¹ Anton Repushko¹

¹Revolut Research ²NVIDIA

Modern financial systems generate vast quantities of transactional and event-level data that encode rich economic signals. This paper presents PRAGMA, a family of foundation models for multi-source banking event sequences. Our approach pre-trains a Transformer-based architecture with masked modelling on a large-scale, heterogeneous banking event corpus using a self-supervised objective tailored to the discrete, variable-length nature of financial records. The resulting model supports a wide range of downstream tasks such as credit scoring, fraud detection, and lifetime value prediction: strong performance can be achieved by training a simple linear model on top of the extracted embeddings and can be further improved with lightweight fine-tuning. Through extensive evaluation on downstream tasks, we demonstrate that PRAGMA achieves superior performance across multiple domains directly from raw event sequences, providing a general-purpose representation layer for financial applications.

Disclaimer: We report only relative improvements, as absolute metrics are commercially sensitive. All examples are synthetic and not from real production data.





Marc Stampfli, Business Director Switzerland
E-Mail: mstampfli@nvidia.com
Linkedin: <https://www.linkedin.com/in/marcstampfli>

